1

Necessity of Research and Development of the Structural Genome Analysis Technology

JUNKO SHIMADA AND SHIN-ICHI MOGI
Life Science and Medical Research Unit

1 Introduction

Completion of the Human Genome Project was announced on April 14, 2003, as the whole human genome was sequenced. The Human Genome Project was proposed in the mid-1980s aiming to sequence the whole human genome. Yet, its realization was thought to be almost impossible since a considerable amount of time was required to complete the whole genome sequencing with the analysis technology at that time. Later on, however, the human genome research advanced rapidly and the international Human Genome Project started in 1990 with the U.S. playing the leading role. When the project was launched, it was planned to sequence the whole human genome of about 3 Gbp and identify all of the 30,000 genes by 2005. However, early completion of the whole genome sequencing was proclaimed in April 2003, mainly because of the great improvements in analyzers[1].

Some people regard the structural genome analysis technology as fully matured because it has developed dramatically through the Human Genome Project in recent years and sequencers have become widely used. Yet, in this era of post genome research, "high-speed and low-cost analysis" is further required. It is thought that the application of conventional technology is insufficient to meet such requirements and the invention of new technology is necessary.

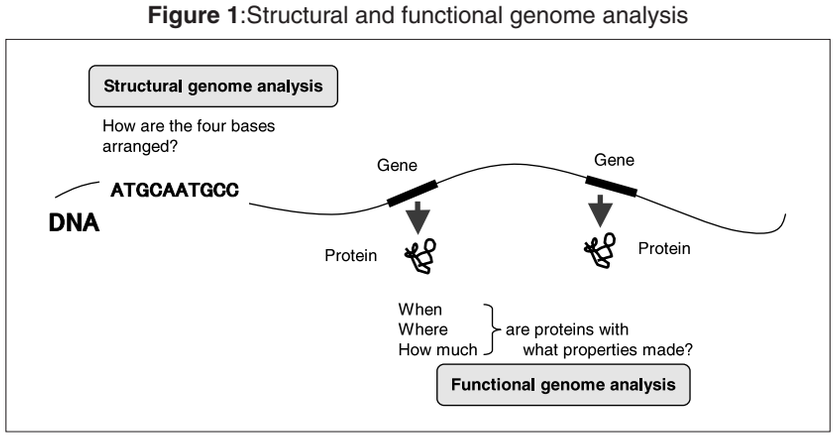
In this report, we will discuss the current status of research and development of structural genome analysis technology and the necessity for their promotion.

2 Structural and functional genome analyses

Genome analysis comprises of structural and functional studies.

In structural genome analysis, the sequence of DNA's four bases is clarified. The genome is the total of the genetic information. The human genome, for example, is contained in the 24 chromosomes consisting of 22 autosomes and two sex chromosomes. DNA carries the genetic information and has four bases of adenine (A), guanine (G), cytosine (C) and thymine (T). The total size of the human genome is about 3 Gbp. DNA sequences vary among individuals and this

Figure 1:Structural and functional genome analysis



phenomenon is called genetic polymorphism. Structural genome analysis has been carried out by sequencing the DNA from an end and comparing the unknown sequence with already known sequences.

In contrast, in functional genome analysis, genome sequences are correlated with phenotypes. In other words, the gene that has the information for producing a protein and the domains that control the production of the protein are identified and the function of the protein is characterized.

3 Development of the structural genome analysis technology

Structural genome analysis has been carried out by sequencing the DNA from an end (DNA sequencing technology) and comparing the sequence of a DNA sample with already known sequences fixed on a substrate based on the binding activity of the hybridized DNAs.

Frederick Sanger disclosed the Sanger method*1 and Allan Maxam and Walter Gilbert disclosed the Maxam-Gilbert method almost at the same time in 1977, as the principles for sequencing the single-strand DNA. At that time, DNA sequencing had been carried out manually using radiation detection, so a researcher could sequence 1,000 bp at most in a year[2]. Since the Sanger method was more suitable for automation than the Maxam-Gilbert method, the Sanger method came to be used widely.

In 1982, Akiyoshi Wada of the University of Tokyo proposed the development of the automated sequencer and invention of element technologies for its realization started. In 1986, Leroy Hood, Lloyd Smith and colleagues disclosed the first automated sequencer. Then, in 1987, Applied Biosystems Inc. marketed the first automated sequencer with Hood's principle applied. Furthermore, in 1990, sequencers based on capillary electrophoresis were developed by three groups led by Lloyd Smith, Barry Karger and Norman Dovichi, respectively. Meanwhile, from 1992 to 1993, Richard Mathies and a group led by Hideki Kambara of Hitachi Ltd. unveiled sequencers based on the capillary-array electrophoresis. Molecular Dynamics and PE

Table 1: Chronology of genome-related technology

1953	Discovery of the double helical structure	J. Watson and F. Crick
1972	Invention of the DNA recombination technology	P. Berg and S. Cohen
1977	Invention of the DNA sequencing method (Sanger method, Maxam-Gilbert method)	F. Sanger, A. Maxam, and W. Gilbert
1980	Proposal of genome mapping by restriction fragment length polymorphism (RFLP)	D. Botstein, R. Davis, M. Skolnick, R. White
1982	Proposal of the automated sequencing system	A. Wada
1984	Development of the pulsed field gel electrophoresis	C. Cantor, D. Schvartz
1985	Invention of the polymerase chain reaction (PCR)	K. Mullis
1986	Development of the autosequencer Invention of a heat-resistant enzyme for PCR	L. Hood, L. Smith, Mullis, K. Saiki
1987	Development of the yeast artificial chromosome (YAC) Autosequencers marketed	D. Burke, M. Olson, G. Carle Applied Biosystems inc.
1989	Development of mapping with the sequence tagged site (STS)	Olson, Hood, Botstein, Cantor
1990	Invention of the capillary electrophoresis	Karger, Smith, N. Dovichi
1992	Development of the bacterial artificial chromosome (BAC)	M. Simon
1993	Development of the capillary array electrophoresis	H. Kambara
1995	Invention of the DNA microarray	P. Brown
1996	DNA chips marketed	Affymetrix
1997	Capillary DNA sequencers marketed	Molecular Dynamics
1998	Capillary DNA sequencers marketed	PE Biosystems Inc.

Source: Authors' compilation on the basis of information provided by Prof. Yoshinobu Baba of the Faculty of Pharmaceutical Sciences at the University of Tokushima and the reference[3].

Biosystems Inc. brought capillary sequencing machines to market in 1997 and 1998, respectively[3]. Capillary sequencing machines are the most widely used among sequencers today.

Assay, detection and analysis of a sample are automated in the capillary sequencing machine. Since electrophoresis, which is the sample assaying process, is conducted in capillaries, high speed and an advanced separation property are obtained. We can sequence as much as 748,800 bp a day with today's capillary sequencing machine (The Applied Biosystems 3730 DNA Analyzer is equipped with 48 capillaries. According to its standard protocol, about 650 bp can be sequenced with a capillary at a time and an hour is required for this process)[4]. In this way, the sequencing technology has been innovated based on the Sanger method and analysis capacity has been boosted.

In the meantime, in order to detect single nucleotide polymorphism (SNP) and screen gene expression, we use the microarray and the DNA chip*2, in which the sample is hybridized with the DNA fixed on the substrate, the difference in sequence between the sample and the fixed DNA is detected, and the sequence of the sample is analyzed. We can also directly analyze the sequences that genetically vary between individuals such as SNP by decoding the sequence one by one from an end and comparing the sequence with a known sequence. Yet, such analysis can be done much easier with the microarray or the DNA chip: We can analyze a number of SNPs at the same time with these methods. However, they are still imperfect technologies and there are hurdles to be overcome such as the noise caused by unspecific adsorption to the substrate.

In connection with this, Patrick Brown and colleagues published in 1995 the first paper on a microarray in which cDNA probes are applied to a glass plate, and Affymetrix produced DNA chips for commercial use in 1996[3].

4 Structural genome analysis technology in post genome research

As the Human Genome Project was completed and post genome researches have been launched, the object of study has developed from structural genome analysis to clarification of the entire biological phenomena and applications of genome studies in medicine and pharmacy.

As the next step to structural genome analysis, researchers are pushing forward functional genome analysis, in which the position and function of genes are characterized based on the DNA sequence, biomolecular studies, such as characterization of protein structures and functions and research on sugar chains, and studies at the cell, tissue and individual levels. We still need to analyze the structure of related genes in such studies. Moreover, besides researches on individual molecules, we need to pursue comprehensive studies such as transcriptome analysis, in which the whole mRNAs existing in a cell at a certain stage are analyzed, and proteome analysis, in which production of all proteins in a cell is studied. Technology for realizing highly efficient analysis is essential in these studies.

Also, researchers are promoting research and development for boosting functional analysis based on decoded genome sequences and for exploiting its fruits. For instance, the difference in genome sequences between individuals is said to be a genetic factor related to the incidence of diseases and drug sensitivity. Researchers are pursuing the development of new ways of diagnosis, treatment and prevention of diseases by accumulating information on genes involved in complex diseases such as cancers and lifestyle-related diseases and by clarifying the mechanism causing their development. Furthermore, researchers are striving to invent new pharmaceuticals by unveiling the characteristics of genes related to drug sensitivity.

In order to achieve such objects, it is necessary to statistically compare genome sequences of individuals and search for genes related to diseases and drug sensitivity. Therefore, we

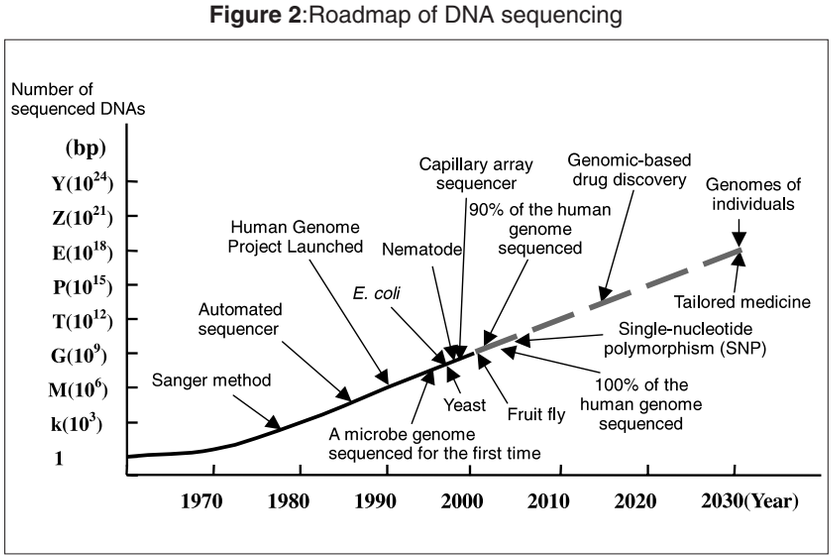
need to analyze statistically a number of human genome sequences. We have only clarified a single set of the human genome in the Human Genome Project. As shown in Figure 2, although the number of sequences registered in the DNA Data Bank of Japan (DDBJ), which is a major international DNA database, has increased exponentially from the 1970s up to now, it has only reached 30 Gbp as of March 2003, including not only human genomes but also genomes of all other organisms like Escherichia coli and nematodes[5]. Since a set of human genome is 3 Gbp, we understand well that we need massive analysis for statistical use, and it is crucial to invent technology for analyzing structures of huge amounts of genomes more efficiently than in conventional methods.

Furthermore, lowering the analysis cost is essential in order to realize medical treatment using genomic information and to employ it in clinical use and for individual treatments as genomic-based testing tools, for example.

In the meantime, there are various fields other than medicine and pharmacy that may benefit from the development of structural genome analysis technology.

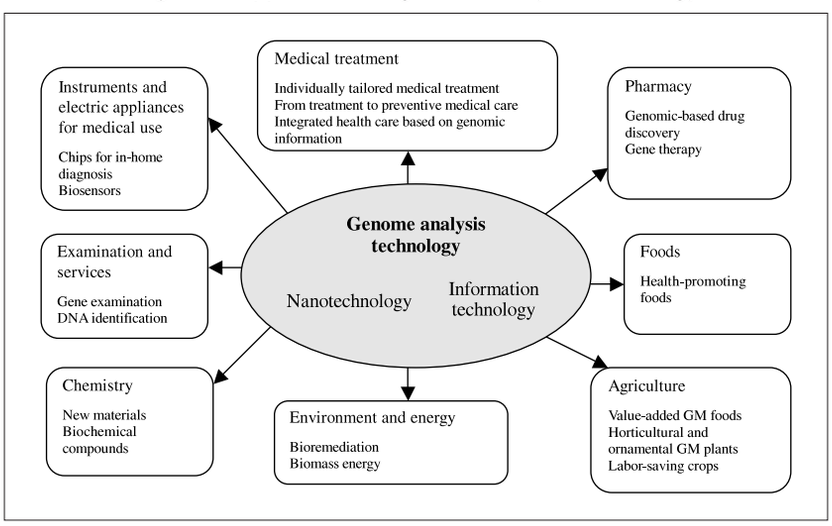
Breeding of plants and animals based on genomic information and invention of health-promoting foods are being pursued in the agriculture and food industry. The

Figure 2:Roadmap of DNA sequencing



Source: Provided by Prof. Yoshinobu Baba of the Faculty of Pharmaceutical Sciences at the University of Tokushima.

Figure 3: Applications of genome analysis technology



Source: Partly Compiled by the authors on the basis of information provided by Prof. Yoshinobu Baba of the Faculty of Pharmaceutical Sciences at the University of Tokushima.

genetically modified crops popularized so far bring such advantages as decreasing the amount of agrichemicals, alleviating farm work and increasing harvest with herbicide and vermin resistance added. Today, researchers are accelerating research and development of crops with high productivity, resistance to soil stresses and nutritional value. Meanwhile, it is hoped that microbe genomes will be used in medicine, pharmacy, chemical production and processing, and environmental conservation. Functional analysis of microbes using their genomic information is promoted to exploit microbe genomes in chemical production and processing and bioremediation, i.e., removing pollutants and restoring the environment using microbes.

Accordingly, highly efficient structural genome analysis will also greatly contribute to research and development in such fields.

5 Recent studies on DNA sequencing technology

Technological innovation in DNA sequencing has been made on the basis of the Sanger method, achieving the invention of the capillary DNA sequencer that is widely used today. Yet, research and development of technologies with principles other than the Sanger method applied are also advancing. Although such technologies are still in the experimental stage, some may be put into practical use. In this section, we will introduce to you some of these new technologies.

5.1 Technologies with the Sanger method applied

Sequencing by mass spectrometry and microchips have been reported as technologies for highly efficient analysis using the Sanger method, as its basic sequencing principle with the analyzers devised.

•Sequencing by mass spectrometry

A mass spectrometer can characterize the mass of molecules precisely and quickly. Ionized molecules are introduced into a vacuum space with a magnetic field and the mass of the molecules is calculated from the drifting time of the ionized molecules. This method is called time-of-flight mass spectrometry (TOF-MS). Mass spectrometry of proteins and DNA molecules became practicable as the electrospray ionization (ESI) and the matrix assisted laser desorption ionization (MALDI) were invented in the 1980s and ionization of these molecules became possible. Today, MALDI-TOF-MS is often used in DNA sequencing. While the principle of sequencing is based on the conventional Sanger method, separation of DNA fragments is conducted by the mass spectrometer. In this method, separation, detection and analysis of DNA fragments are completed on the second time scale, and, moreover, only a few hundred nanoliters of the sample are required. Yet, so far, only about 100 bp can be analyzed in one course of analysis[2].

•Sequencing by microchips

In sequencing by microchips, a microchip electrophoresis device in which minute channels are formed on a glass or plastic substrate is used. This sequencing technology has the same principle as the capillary electrophoresis. Since the microchannels have very small heat capacity, high voltage can be applied. Thus, speedy analysis with high separation performance can be conducted, and little amounts of samples are necessary for analysis. In addition, semiconductor technology is applied in the production of this chip, so the experimental system can be parallelized through integration with sample preparation and detection carried out on one chip. This concept is called “micro total analysis system (μTAS)” or “lab-on-a-chip”[2]. Today, the experimental course of amplifying DNAs by the polymerase chain reaction (PCR) and separating its products according to their molecular weights with electrophoresis can be conducted on several samples in parallel on one chip. However, a chip on which DNA can be sequenced has not yet been invented.

5.2 Technologies with other principles applied

Researchers are striving to sequence DNAs directly without using the Sanger method as in microscopic sequencing and nanopore sequencing. Also, DNA sequencing technology using DNA chips has been proposed.

•Microscopic sequencing

Direct observation of DNAs has become possible along with the rapid improvement of electron microscopes and scanning probe microscopes (SPM). Some research groups are endeavoring to sequence DNAs by directly observing bases on the DNA chains, and there are reports on the observation of the DNA double helix and distinction between adenine and thymine[2].

•Nanopore sequencing

In nanopore sequencing, DNA is sequenced when it passes through a nanopore, or a nano-level pore, with a diameter slightly larger than the molecular diameter of the DNA. When the DNA molecule enters the pore, the electrical property of the nanopore changes depending on the DNA bases passing. Researchers are attempting to sequence DNAs using this base dependency of the electric current. Since a DNA molecule passes a nanopore in several milliseconds, rapid sequencing can be realized if this nanopore sequencing becomes feasible. So far, there has been a report describing the difference in change in electrical property between polyadenine and polycytosine[2].

•Sequencing with DNA chips

In this method, all kinds of oligonucleotides of a specific length are fixed on a DNA chip, and the sample to be analyzed is fluorescence-labeled and applied to the chip. The oligonucleotides hybridized with the sample indicate that their sequences exist in the sample. Accordingly, the sequences of all hybridized oligonucleotides are compared and the whole sequence of the sample is estimated based on the overlaps of the oligonucleotide sequences. However, there is a disadvantage that only a sample with length corresponding to the square root of the number of oligonucleotides fixed on the substrate can be sequenced. For example, there are 65,536 kinds of oligonucleotides with a length of 8 bp, and even if a chip contains all these oligonucleotides, the maximum length that can be sequenced is only 256 bp. When a longer sample is sequenced, far more various kinds of sequences must be loaded on the chip. There is a report that sequencing with DNA chips may be put into practical use if a large variety of sequences can be fixed on the chip and hybridization can be detected electronically[6].

5.3 Technology supporting DNA sequencing

Since sequencers today can deal with only a limited length of samples, we need to fragment DNA when we analyze long sequences like genomes. However, fragmentation is carried out *in vitro* and the order of the DNA fragments becomes random. Thus, after sequencing each fragment, we need to ascertain overlaps between fragments by computer analysis and rearrange them. This step takes time and tends to cause mistakes. As a method to overcome such disadvantages, single-molecule DNA sequencing has been proposed.

•Single-molecule DNA sequencing

Researchers try to realize single-molecule DNA sequencing, in which a single DNA molecule, i.e., one DNA strand, is physically fixed linearly and fragmented from an end one by one, and the fragments are amplified by PCR and sequenced with the Sanger method. This procedure is favorable in that we do not need to rearrange the DNA fragments after sequencing, because the fragments are aligned according to the order in the original sample. There is a report that a sample was successfully extended, fixed, fragmented and isolated. Another report says a sample could be amplified with PCR from a single molecule only[2].

6 Conclusion

Development of the genome analysis technology has been emphasized since the start of the Human Genome Project. For example, in the U.S. Human Genome Project, certain numerical targets for analysis speed and costs were set to boost genome sequencing capacity. Also in Japan, many government proposals for promoting human genome analysis since 1987 have stressed that the nation should advance DNA analysis technology.

As a matter of fact, the Special Coordination Funds for Promoting Science and Technology of the former Science and Technology Agency

supported "Research on development of DNA extraction, analysis and synthesis technology" from 1981 to 1983 and "Research for development of common basic technology for cancer studies" from 1984 to 1989. Moreover, RIKEN promoted "Development of the Human Genome Analyzer (HUGA)" from 1987 to 1994.

It is notable that the concept of the autosequencer was proposed by a Japanese researcher. Japan has obtained and exploited fruits of research and development of element technologies. However, in reality, most of the sequencers that Japanese laboratories purchase are produced by foreign companies. For example, the market share of the capillary DNA sequencers of foreign companies in Japan was 99% by value in FY2001[7].

Forefront research starts from device invention, and innovative measurement technologies activate new spheres of study. Thus, if we have domestic bases for developing state-of-the-art analyzers, we can enhance the quality of research and development more and more. Structural genome analysis technology contributes to improvements in medicine, pharmacy, agriculture, the food industry, chemistry, environmental studies and so forth. Therefore, progress in structural genome analysis technology will elevate the standard of research and development of many studies.

Recently, the Japanese government announced new programs for developing analyzers. The Biotechnology Strategy Guidelines adopted on December 6, 2002 by the Biotechnology Strategy Council of the Japanese Cabinet, which aims at applying and industrializing fruits of biotechnological research, enhancing the nation's quality of life and bolstering industrial competitiveness, advocates the promotion of cooperation with information technology and nanotechnology and intensified investments to the development of biotechnological devices. Moreover, the policy for distributing resources such as budgets and talent in science and technology in FY 2004 adopted by the Council for Science and Technology Policy, Cabinet Office on June 19, 2003 sets development of forefront technologies and devices for analysis of genes and proteins as a prioritized target in life science. In addition, the Ministry of Education, Culture, Sports, Science and Technology organized a panel on the development of forefront technologies of measurement and analysis in June 2003, and is scrutinizing how to practically promote research and development of devices.

In such a current, we need to encourage further research and development with attention to several points as follows.

As we enter the post genome era and research objects will expand from structural genome analysis into functional genome analysis and clarification of the relationship between genes and diseases for medical applications, genome analysis technology is expected to grow vigorously. Taking this circumstance into consideration, we need to continuously back up basic research, unveiling new principles and applied research to exploit the fruits of basic studies.

Furthermore, we should support technologies in the development stage until devices applicable in laboratories and medical treatment are invented. We must develop such devices as systems including reagents and analysis software.

Acknowledgements

This report has been compiled based on "Trends and Prospects in Next Generation Nanodevice Research," a lecture by Dr. Yoshinobu Baba, professor of the Faculty of Pharmaceutical Sciences at the University of Tokushima and chief of the Single-molecule Bioanalysis Laboratory of the National Institute of Advanced Industrial Science and Technology, held on May 12, 2003 at the National Institute of Science and Technology Policy and by adding the results of our investigations. Our sincere gratitude to Prof. Baba, who provided us with invaluable suggestions and information.

Glossary

*1 Sanger method

A method for sequencing a single strand DNA. When a polynucleotide chain with a sequence complementing a single strand DNA is synthesized with enzymes, the synthesis can be stopped artificially at a proper position of the nucleotides. That is,

although deoxyribonucleoside triphosphates (dNTPs) serve as the substrate for synthesizing complementary sequences, a small amount of dideoxyribonucleoside triphosphates (ddNTPs) is added to dNTPs. Enzymes incorporate dNTPs and ddNTPs into the DNA strand without distinguishing between them. When a ddNTP is incorporated, DNA synthesis stops. The polynucleotides produced are electrophoresed, separated according to their molecular weights, and detected.

*2 Microarray and DNA chip[6]

In a microarray and DNA chip, particular DNAs are aligned at high density on a substrate of glass or polymers. In the microarray, DNAs are dripped onto the substrate. On the other hand, in the DNA chip, oligonucleotides are synthesized on the surface of the chip and DNAs can be aligned at higher density than in the microarray. When the DNA sample to be analyzed is fluorescence-labeled and reacts with the DNAs on the substrate, hybridization occurs if there are complementary sequences of the sample. Hybridization between the sample and the DNAs on the substrate is detected with a fluorescence detector and analyzed with a computer.

References

[1] Homepage of the Human Genome Project of the U.S. Department of Energy, : http://www.ornl.gov/TechResources/Human_Genome/home.html

[2] Hirano, Ken and Baba, Yoshinobu, "Next Generation DNA Sequencer," Bio Venture 2002, 2 (3): 38-44 (in Japanese).

[3] Roberts et al., "A History of the Human Genome Project," Science 2001, 291 (5507): 1195.

[4] Homepage of Applied Biosystems Japan Ltd.,: http://www.appliedbiosystems.co.jp/website/jp/home/index_g.jsp

[5] Homepage of DDBJ, :http://www.ddbj.nig.ac.jp/Welcome-j.html

[6] Brown, T. A., "Genomes 2 (Japanese version)," Medical Science International, 2003.

[7] R&D Corp., "Scientific Instruments Almanac," , 2002 (in japanese).

(Original Japanese version: published in August 2003)